# Supplementary Material for "Bayesian Double Feature Allocation for Phenotyping with Electronic Health Records"

## A. Additional Results for Simulation Studies

**Sensitivity analysis.** We chose $\tau = \tau_w^2 = 100$ in our simulation studies and application. Here we perform a sensitivity analysis on a range of hyperparameter values $(\tau, \tau_w^2) \in \{(50, 100), (75, 100), (125, 100), (150, 100), (100, 50), (100, 75), (100, 125), (100, 150)\}$ using simulations, each repeated 50 times. We denote $e_K$ the proportion of the repetitions where $\widehat{K} \neq K$, $e_A$ the mis-allocation rate, and $e_B$ and $e_C$ the error rates in estimating $\boldsymbol{B}$ and $\boldsymbol{C}$. These four operating characteristics are summarized in Table 1. The proposed model is not sensitive to the choice of hyperparameters.

**Hierarchical prior.** We test the proposed method with the hierarchical prior in (2) using $a_\sigma = b_\sigma = 0.01$. The performance is very similar to that of the non-hierarchical prior: $e_K = 4\%$, $e_A = 0.03$, $e_B = 0.01$ and $e_C = 0.01$.

**Consensus Monte Carlo.** We propose a simple consensus Monte Carlo (CMC) algorithm to scale up DFA to large sample size. The CMC algorithm goes as follows.

1. Randomly split the data into $S$ subsets where $S$ depends on the computation resources at hand.
2. Run the MCMC algorithm described in Section 4 to each subset in parallel.
3. Combine the posterior draws of two latent diseases if they share a similar set of symptoms. Specifically, we consider two latent diseases to be similar if the proportion of different symptoms is less than $\epsilon$.

CMC essentially divides a large dataset into many smaller subsets which are easier and faster to process. We assess the performance as well as the speed of the proposed CMC with simulations. We follow the same simulation setup in Scenario I except that we now expand the sample size 50 times larger $n = 300 \times 50 = 15,000$. We choose $S = 50$ and $\epsilon = 0.50$. The total computation time is less than 5 minutes. Though it overestimates $K$ by 2, it estimates $\boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{C}$ quite well (after removing the two extra columns): $e_A = 0.03$, $e_B = 0.01$ and $e_C = 0.01$. The performance is relatively robust with respect to the choice of $\epsilon$: $\epsilon = 0.25$ and $\epsilon = 0.75$ yield virtually the same performance as $\epsilon = 0.50$. However, when we set $\epsilon$ to 0.05, the estimated number of latent diseases goes up to 67 whereas the truth is 6.

Table 1: Sensitivity analysis of $\tau$ and $\tau_w^2$. We use the following notations. $e_K$: the proportion of the repetitions where $\widehat{K} \neq K$. $e_A$: the mis-allocation rate. $e_B$ and $e_C$ are error rates in estimating $\boldsymbol{B}$ and $\boldsymbol{C}$.

| | $\tau$, $\tau_w^2 = 100$ | | | | $\tau_w^2, \tau = 100$ | | | |
|---|---|---|---|---|---|---|---|---|
| | 50 | 75 | 125 | 150 | 50 | 75 | 125 | 150 |
| $e_K$ | 0.10 | 0.06 | 0.04 | 0.04 | 0.06 | 0.02 | 0.04 | 0.06 |
| $e_A$ | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 |
| $e_B$ | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 |
| $e_C$ | 0.03 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

# B. Additional Results for EHR Data Analysis

**Parameter estimation.** We report the posterior mean of $\{\boldsymbol{W}_j, \zeta_j\}_{j=1}^p$ and $\{\boldsymbol{W}_l^-, \boldsymbol{W}_l^+, \eta_l^-, \eta_l^+\}_{l=1}^q$ on inverse-logit scale in Figures 1-3. Inverse-logit is defined as

$$\text{inv-logit}(x) = \frac{\exp(x)}{1+\exp(x)}.$$

Each dot is a symptom-disease relationship of which the weight is marked on the y-axis. Dots of the same type correspond to the same symptom of which the baseline weight is marked on the x-axis.

**MCMC convergence.** The acceptance rates for parameters updated by Metropolis-Hasting transition probabilities are 25% ($\{\zeta_j\}_{j=1}^p$), 11% ($\{\eta_l^-, \eta_l^+\}_{l=1}^q$) and 33% ($\{\boldsymbol{W}_j\}_{j=1}^p$, $\{\boldsymbol{W}_l^-, \boldsymbol{W}_l^+\}_{l=1}^q$). We monitor the MCMC convergence using the potential scale reduction factor (PSRF Gelman & Rubin 1992). Based on five parallel chains with different starting values, we calculate PSRF for the sampling log-density (defined in Section 2.3), a quantity that is invariant to "label switching". The point estimate and the upper 95% confidence limit of PSRF are both equal to 1, which indicates no lack of convergence.

**Goodness-of-fit diagnostics.** We perform a chi-square test proposed by Yuan & Johnson (2012). We follow their procedure.
1. Let $f_z$ and $F_z$ be the probability mass function and cumulative distribution function for $z_{ij}$. Define $z_{ij}^{(t)} = F_z(z_{ij}-1|\boldsymbol{\theta}^{(t)}) + u_{ijt} f_z(z_{ij}|\boldsymbol{\theta}^{(t)})$ where $u_{ijt} \sim Unif(0,1)$ and $\boldsymbol{\theta}^{(t)}$ is the $t$th Monte Carlo sample of model parameters $\boldsymbol{\theta}$.
2. Define similar transformation for categorical variables $y_{il}$. Note that the cumulative distribution function is not well defined for a categorical variable and therefore we collapse categorical distribution into Bernoulli distribution. Specifically, let $\tilde{y}_{il} = I(y_{il} < 0)$ and let $f_y$ and $F_y$ be the probability mass function and cumulative distribution function for $\tilde{y}_{il}$. Define $y_{il}^{(t)} = F_z(\tilde{y}_{il}-1|\boldsymbol{\theta}^{(t)}) + v_{ilt} f_y(\tilde{y}_{il}|\boldsymbol{\theta}^{(t)})$ where $v_{ilt} \sim Unif(0,1)$. Importantly, $z_{ij}^{(t)}$'s and $y_{il}^{(t)}$'s are iid standard uniform random variables when models are correctly specified.
3. For each $t$, partition the sample space of $z_{ij}^{(t)}$ into $K_1 = 2$ groups based on whether $f_z(z_{ij}|\boldsymbol{\theta}^{(t)})$ is greater than or less than 0.5. Partition the sample space of $y_{il}^{(t)}$ into $K_2 = 2$

groups in the same way. Let $K = K_1 + K_2$ be the total number of groups.
3. Within the $k$th group for $k = 1, \ldots, K$, place the transformed items $z_{ij}^{(t)}$'s and $y_{il}^{(t)}$'s into $L = 10$ bins according to quantiles of standard uniform distribution. Let $n_k^{(t)}$ be the total number of items in group $k$ and let $O_{kl}^{(t)}$ be the observed number of items in bin $l$ and group $k$. Compute the $\chi^2$ statistic

$$d_k^{(t)} = \sum_{l=1}^{L} \left( \frac{O_{kl}^{(t)} - n_k^{(t)}/L}{\sqrt{n_k^{(t)}/L}} \right)^2 .$$

5. Sum the $\chi^2$ statistics to obtain the global pivotal discrepancy measure

$$d^{(t)} = \sum_{k=1}^{K} d_k^{(t)},$$

which is approximately $\chi^2_{K(L-1)}$ distributed when the sample is large and the model is correctly specified.

In Figure 4, we plot the histogram of $d^{(t)}$'s and the reference $\chi^2_{36}$ distribution, which suggests adequate fit.

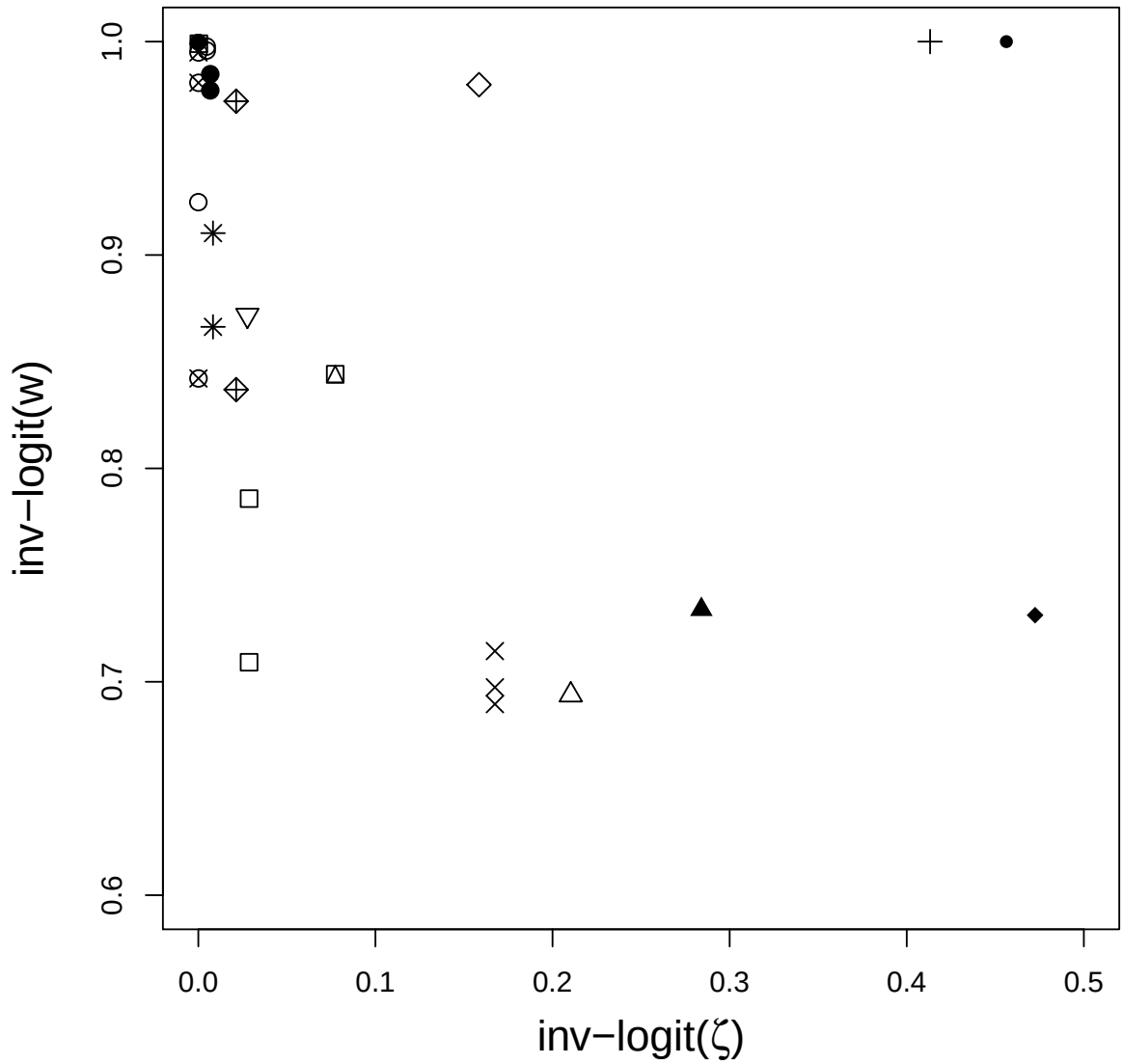


Figure 1: Posterior mean of $\boldsymbol{W}_j$ and $\zeta_j$ on inverse-logit scale for $j = 1, \ldots, p$. Each dot is a symptom-disease relationship of which the weight is marked on the y-axis. Dots of the same type correspond to the same symptom of which the baseline weight is marked on the x-axis.

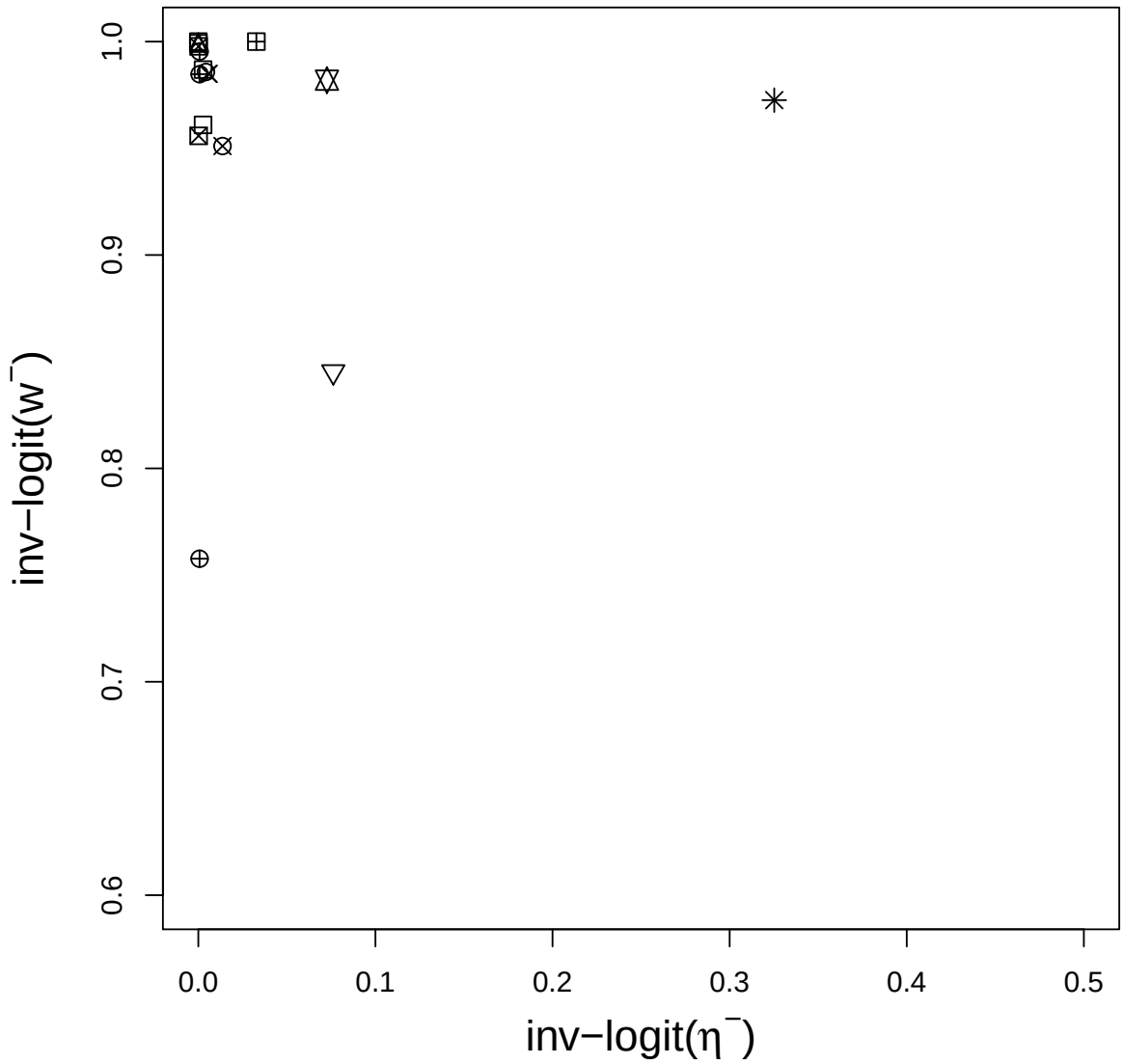


Figure 2: Posterior mean of $\boldsymbol{W}_l^-$ and $\eta_l^-$ on inverse-logit scale for $l = 1, \ldots, q$. Each dot is a symptom-disease relationship of which the weight is marked on the y-axis. Dots of the same type correspond to the same symptom of which the baseline weight is marked on the x-axis.

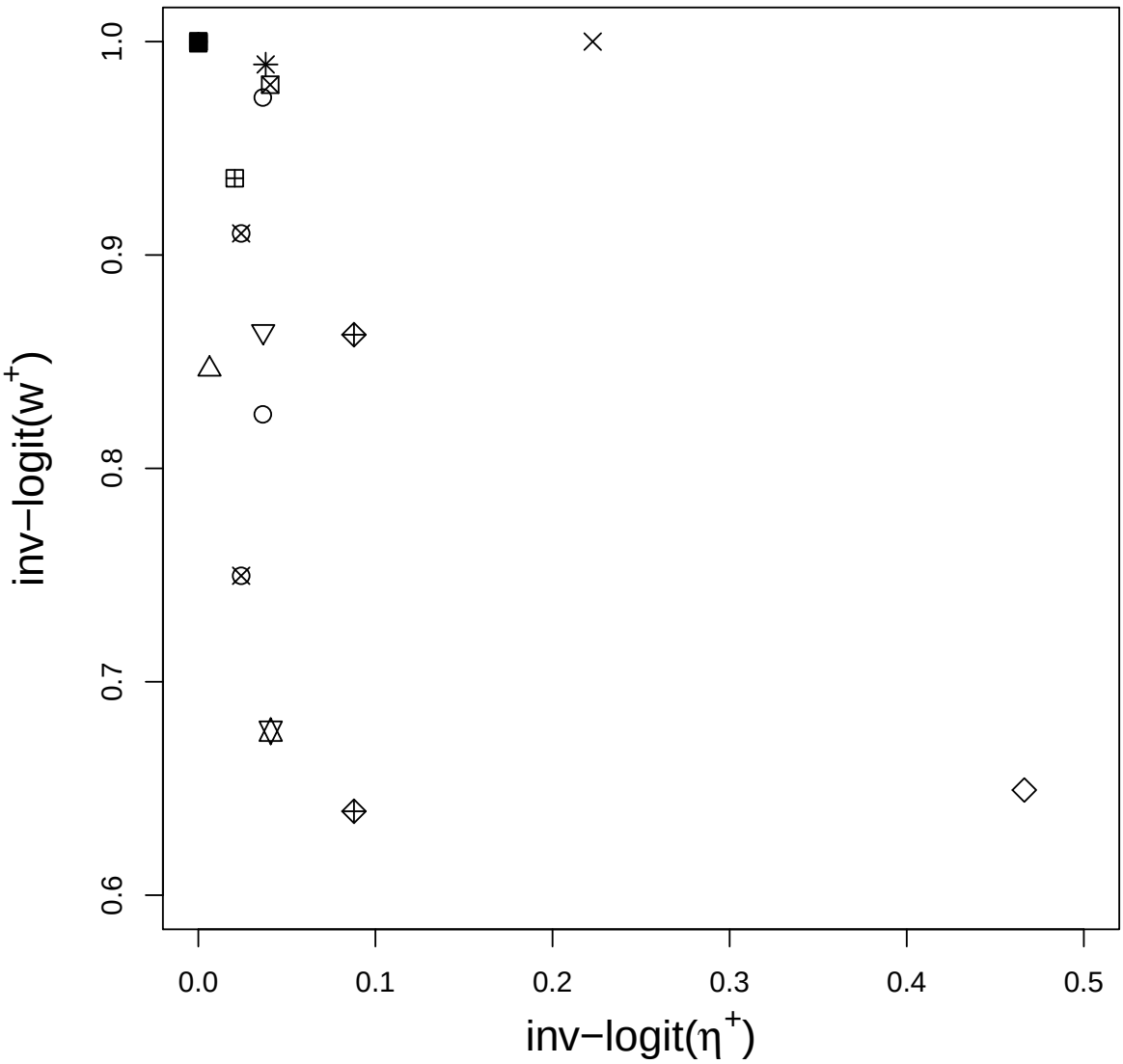


Figure 3: Posterior mean of $\boldsymbol{W}_l^+$ and $\eta_l^+$ on inverse-logit scale for $l = 1, \ldots, q$. Each dot is a symptom-disease relationship of which the weight is marked on the y-axis. Dots of the same type correspond to the same symptom of which the baseline weight is marked on the x-axis.

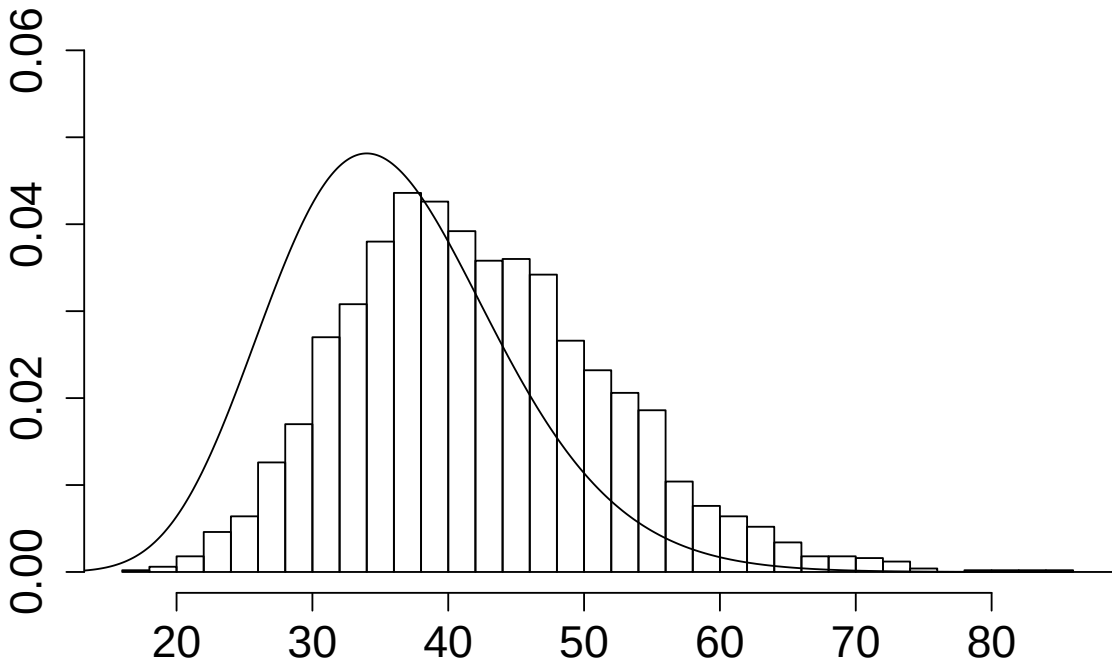


Figure 4: Goodness-of-fit diagnostics. The posterior distribution of the discrepancy measure is shown as histogram. The reference $\chi^2_{36}$ distribution is shown as a density curve.

# References

Gelman, Andrew, & Rubin, Donald B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472.

Yuan, Ying, & Johnson, Valen E. 2012. Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics*, **68**(1), 156–164.